



Group 4



SECB3203-01

PROGRAMMING FOR BIOINFORMATICS

ALZHEIMER'S DISEASE PREDICTION



Lecturer: Dr. Seah Choon Sen





GROUP MEMBERS



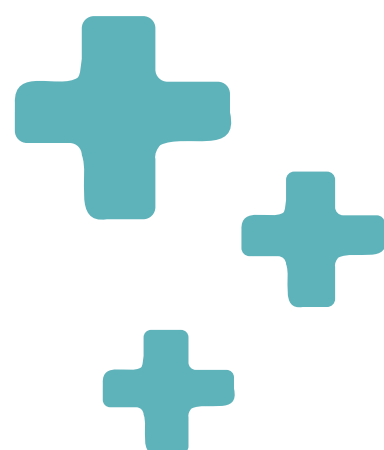
TAN ZHAO HONG
A23CS0188

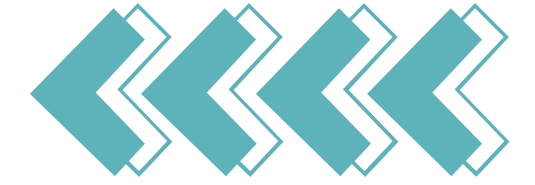


CHIN PEI WEN
A23CS0065



KOO XUAN
A23CS0300





INTRODUCTION



- Alzheimer's Disease (AD) is a progressive neurodegenerative disorder, leading to cognitive decline and memory loss.
- Early diagnosis is critical for timely intervention, treatment planning and improving quality of life.
- Objective: Develop a machine learning model to predict Alzheimer's disease using clinical, demographic, and lifestyle data.



PROBLEM STATEMENT

Current Challenges in AD Diagnosis:

- **Delayed diagnosis** – often at advanced stages.
- **Limited accessibility** – imaging & lab tests not widely available.
- **High costs** – imaging and biomarker testing expensive
- **Subjective assessments** – cognitive tests rely on practitioner judgment.
- **Data integration challenges** – scattered patient records.
- **Lack of predictive tools** – focus is on confirming diagnosis, not risk prediction.

DATASET OVERVIEW



Source: Kaggle Alzheimer's Disease Dataset

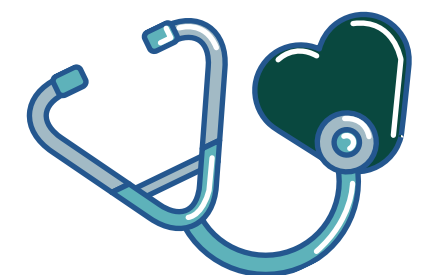
Target variable:

Diagnosis (0 = No Alzheimer's, 1 = Alzheimer's)

Data domains:

- Demographics: Age, Education, Gender, Ethnicity
- Lifestyle factors: Diet, Physical Activity, Sleep, Alcohol, Smoking
- Medical history: Family history of AD, Cardiovascular disease, Diabetes, Hypertension, Depression
- Clinical measurements: BMI, Blood pressure, Cholesterol
- Cognitive assessments: MMSE, ADL, Functional Assessment, Memory complaints
- Symptoms: Behavioral changes, Confusion, Disorientation, Personality changes

Total 35 features



DATA PREPROCESSING



- Handled missing values (median imputation).
- Normalized numerical features.
- Binning & encoding for categorical variables (e.g., age groups).
- Split data into training (80%) and test (20%) sets.

DATA PREPROCESSING



FEATURE REMOVAL

Removed non-medical / weak-evidence features:

- PatientID
- DoctorInCharge
- Gender
- Ethnicity
- AlcoholConsumption
- SleepQuality

Purpose:

- Reduce noise
- Prevent bias
- Improve model focus on medical relevance

```
=====
REMOVING NON-MEDICAL AND WEAK EVIDENCE FEATURES
=====
```

```
Removing features: ['PatientID', 'DoctorInCharge', 'Gender', 'Ethnicity', 'AlcoholConsumption', 'SleepQuality']
New dataset shape: (2149, 29)
```

DATA PREPROCESSING

MISSING VALUES

Checked missing values across all features

Applied median imputation for numerical features

Reason:

- Robust to outliers
- Preserves data distribution

Ensured dataset completeness before modeling

```
# Handle missing values
if X.isnull().sum().sum() > 0:
    print("\n3. Handling missing values...")
    X = X.fillna(X.median())
    print("Missing values after handling:")
    print(X.isnull().sum())
else:
    print("\n3. No missing values detected!")
```

--- Identifying and Handling Missing Values ---

1. Missing Values Count:

Age	0
EducationLevel	0
BMI	0
Smoking	0
PhysicalActivity	0
DietQuality	0
FamilyHistoryAlzheimers	0
CardiovascularDisease	0
Diabetes	0
Depression	0
HeadInjury	0
Hypertension	0
SystolicBP	0
DiastolicBP	0
CholesterolTotal	0
CholesterolLDL	0
CholesterolHDL	0
CholesterolTriglycerides	0
MMSE	0
FunctionalAssessment	0
MemoryComplaints	0
BehavioralProblems	0
ADL	0
Confusion	0
Disorientation	0
PersonalityChanges	0
DifficultyCompletingTasks	0
Forgetfulness	0
dtype: int64	

2. Visualizing Missing Values:

3. No missing values detected!

DATA PREPROCESSING

NORMALIZATION & ENCODING

Min-Max normalization applied
(example: Age feature)

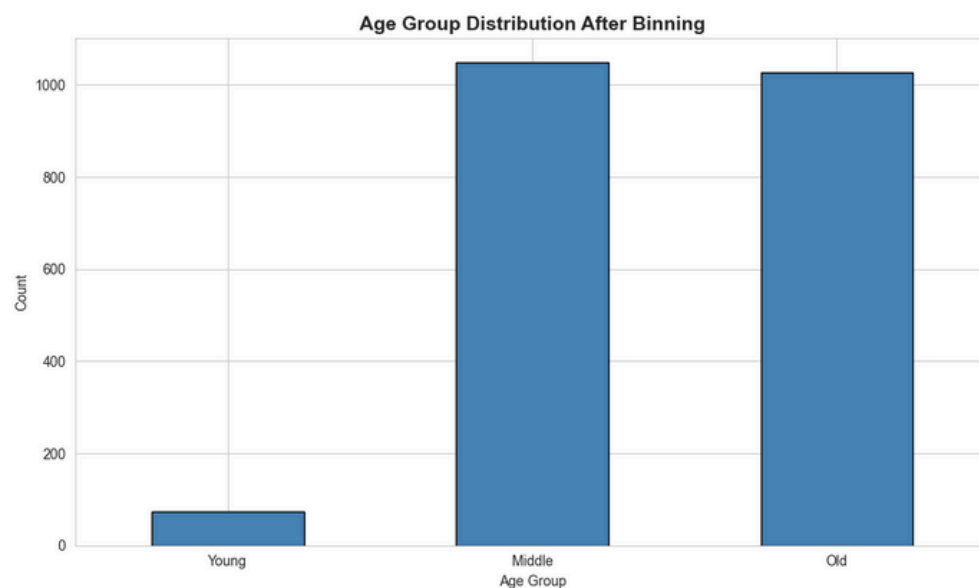
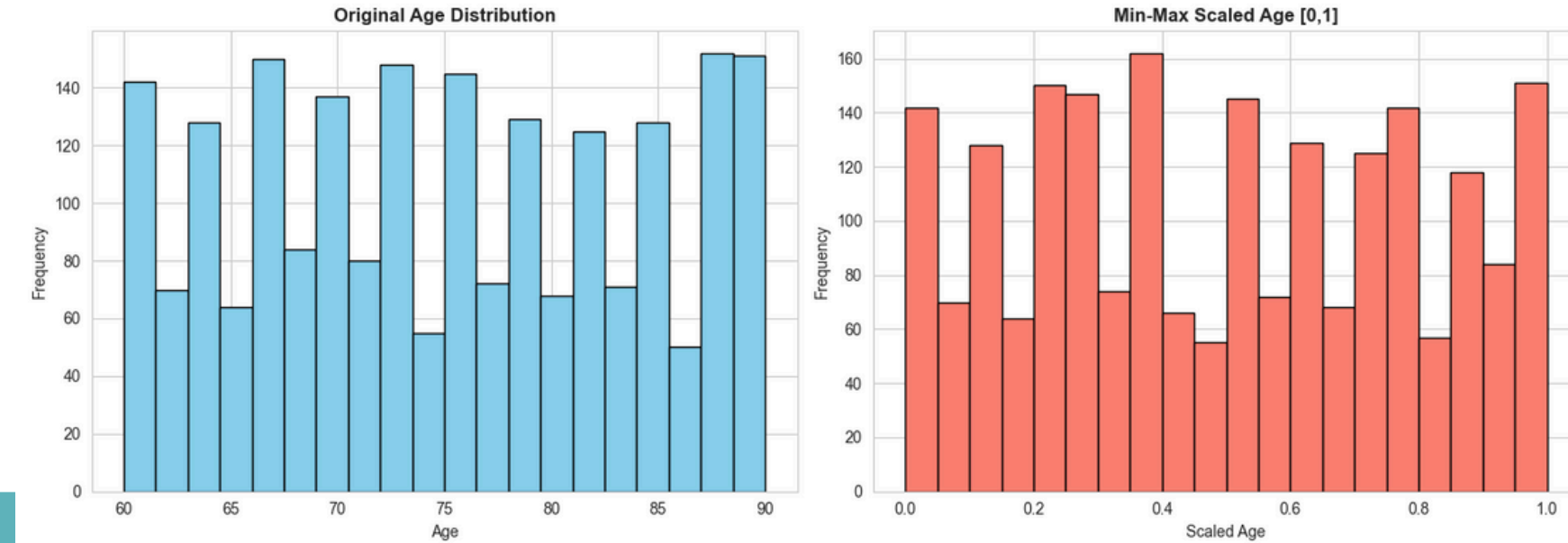
Z-score normalization using StandardScaler

Age binning:

- Young (≤ 60)
- Middle-aged (61–75)
- Old (> 75)

One-hot encoding for categorical variables

Stratified train-test split (80:20)



--- Indicator Variables (One-Hot Encoding) ---

1. Categorical columns to encode: ['Age_binned']

2. Before encoding (first 5 rows):

	Age_binned
0	Middle
1	old
2	Middle
3	Middle
4	old

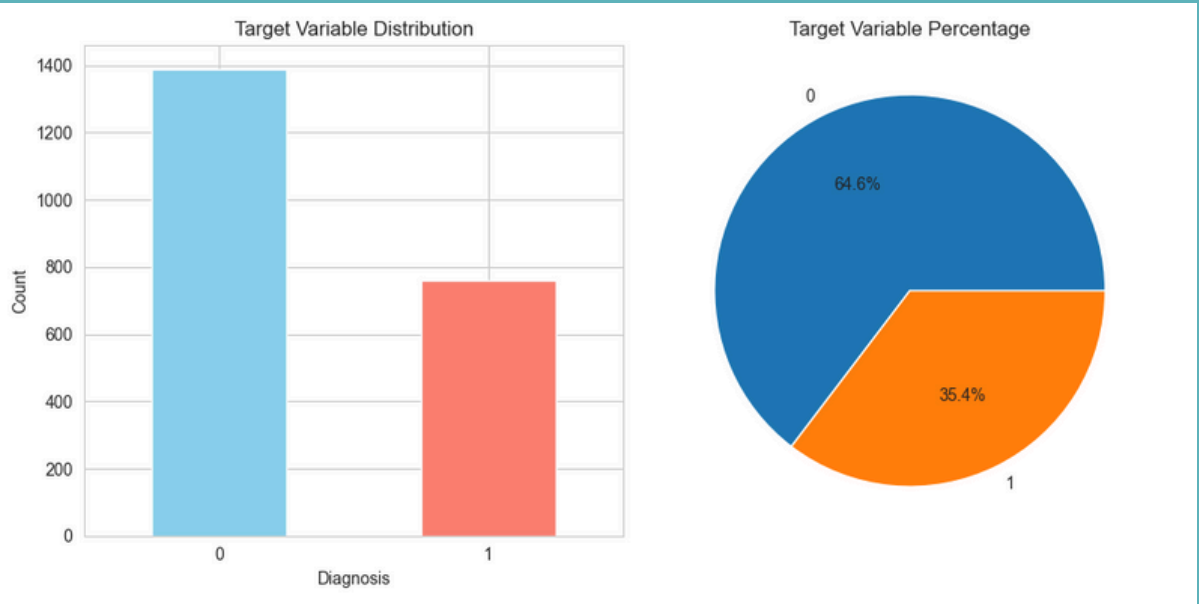
3. After one-hot encoding:

Original features: 1
New indicator columns created: 2
Total columns now: 30

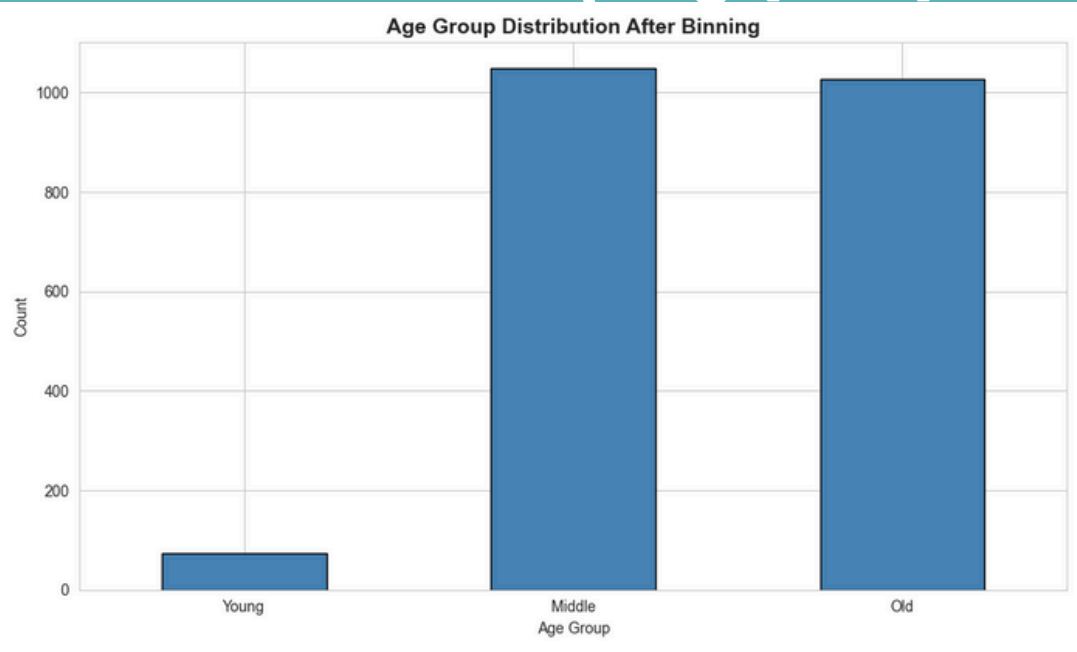
EXPLORATORY DATA ANALYSIS (EDA)



Target variable distribution analysis



Basic of Grouping (AGE)



Descriptive Statistics

[2] Performing Exploratory Data Analysis...

Statistical Summary:

	Age	EducationLevel	BMI	Smoking	PhysicalActivity	...	Disorientation	PersonalityChanges	DifficultyCompletingTasks	Forgetfulness	Diagnosis
count	2149.000000	2149.000000	2149.000000	2149.000000	2149.000000	...	2149.000000	2149.000000	2149.000000	2149.000000	2149.000000
mean	74.908795	1.286645	27.655697	0.288506	4.920202	...	0.158213	0.150768	0.158678	0.301536	0.353653
std	8.990221	0.904527	7.217438	0.453173	2.857191	...	0.365026	0.357906	0.365461	0.459032	0.478214
min	60.000000	0.000000	15.008851	0.000000	0.003616	...	0.000000	0.000000	0.000000	0.000000	0.000000
25%	67.000000	1.000000	21.611408	0.000000	2.570626	...	0.000000	0.000000	0.000000	0.000000	0.000000
50%	75.000000	1.000000	27.823924	0.000000	4.766424	...	0.000000	0.000000	0.000000	0.000000	0.000000
75%	83.000000	2.000000	33.869778	1.000000	7.427899	...	0.000000	0.000000	0.000000	1.000000	1.000000
max	90.000000	3.000000	39.992767	1.000000	9.987429	...	1.000000	1.000000	1.000000	1.000000	1.000000

[8 rows x 29 columns]

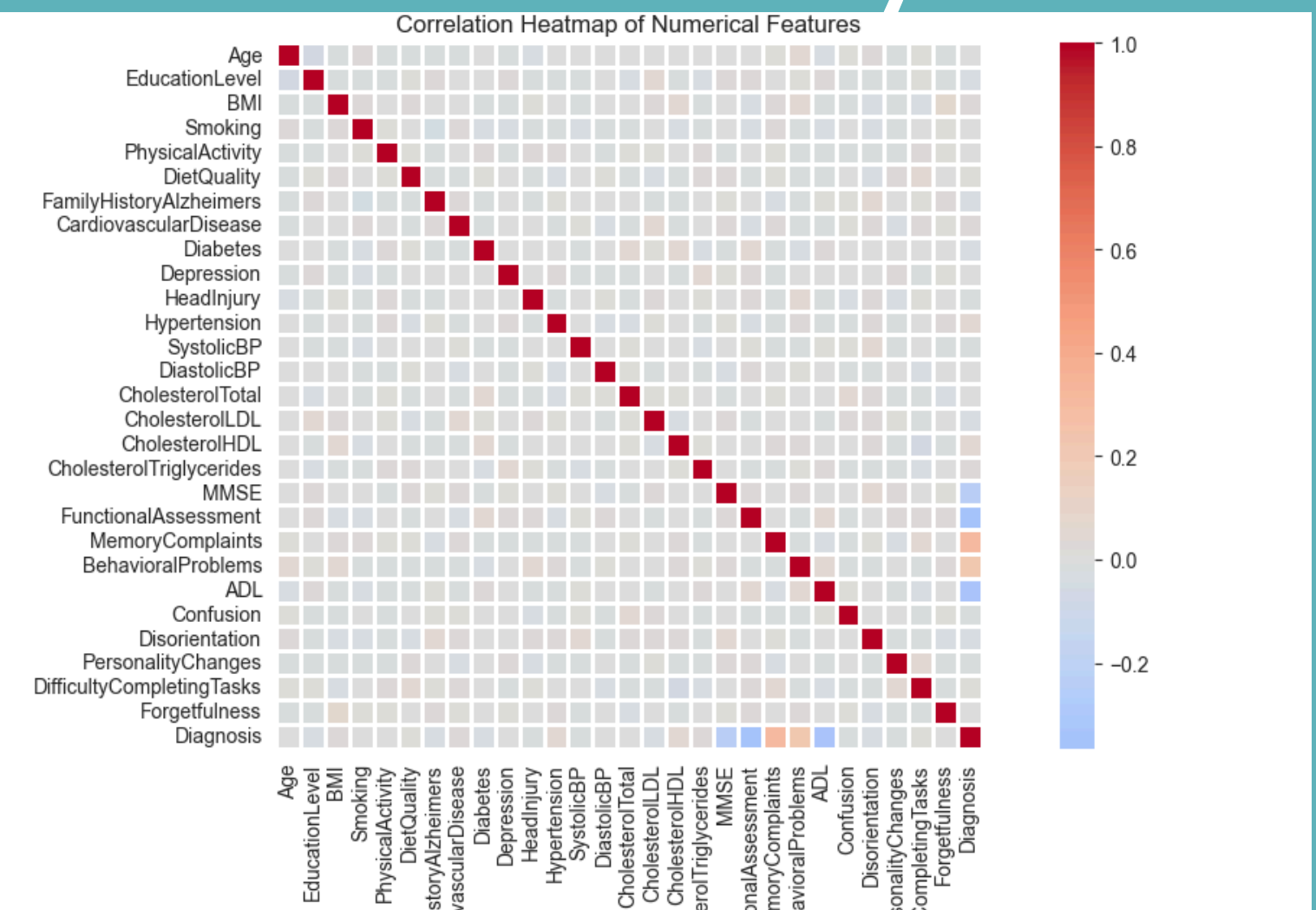


EXPLORATORY DATA ANALYSIS (EDA)

ANOVA test

```
--- ANOVA Test (Age vs Diagnosis) ---  
F-statistic: 0.06467450176025781  
P-value: 0.799279022412292
```

Correlation analysis



MODEL DEVELOPMENT

Alzheimer's disease diagnosis, is binary (0 = No Alzheimer's, 1 = Alzheimer's)

Libraries

```
4 import numpy as np
5 import pandas as pd
6 import matplotlib.pyplot as plt
7 import seaborn as sns
8 from sklearn.model_selection import train_test_split, cross_val_score, GridSearchCV
9 from sklearn.preprocessing import StandardScaler, LabelEncoder
10 from sklearn.metrics import (accuracy_score, precision_score, recall_score,
11                             | f1_score, confusion_matrix, classification_report,
12                             | roc_auc_score, roc_curve)
13 # from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
14 from sklearn.linear_model import LogisticRegression
15 from sklearn.svm import SVC
16 from sklearn.tree import DecisionTreeClassifier
17 from sklearn.neighbors import KNeighborsClassifier
18 import warnings
19 warnings.filterwarnings('ignore')
```

Models implemented:

- Logistic Regression
- Decision Tree
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)



MODEL EVALUATION



MODEL TRAINING AND EVALUATION

Logistic Regression:

Accuracy: 0.8140
Precision: 0.8145
Recall: 0.8140
F1-Score: 0.8142
CV Accuracy: 0.8395 (+/- 0.0211)

Decision Tree:

Accuracy: 0.8884
Precision: 0.8884
Recall: 0.8884
F1-Score: 0.8884
CV Accuracy: 0.9069 (+/- 0.0117)

SVM:

Accuracy: 0.8326
Precision: 0.8305
Recall: 0.8326
F1-Score: 0.8302
CV Accuracy: 0.8336 (+/- 0.0169)

KNN:

Accuracy: 0.7465
Precision: 0.7418
Recall: 0.7465
F1-Score: 0.7308
CV Accuracy: 0.7254 (+/- 0.0135)

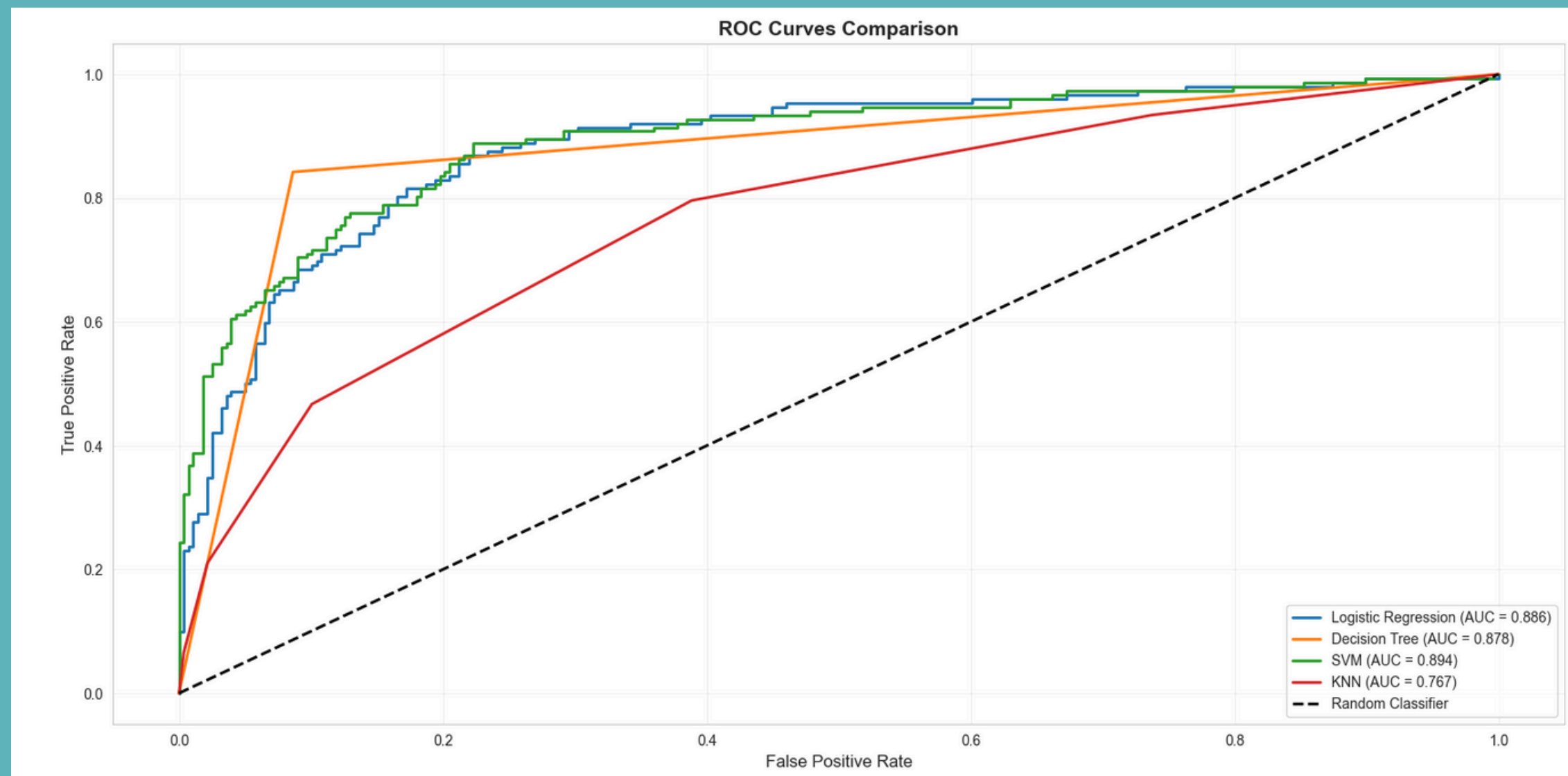
EVALUATION METRICS

- Accuracy
- Precision
- Recall
- F1-score
- Cross-validation accuracy (5-fold)

MODEL EVALUATION



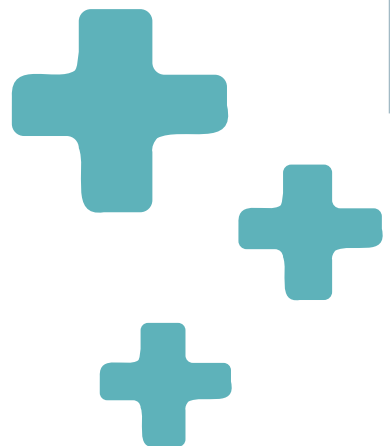
ROC CURVE ANALYSIS





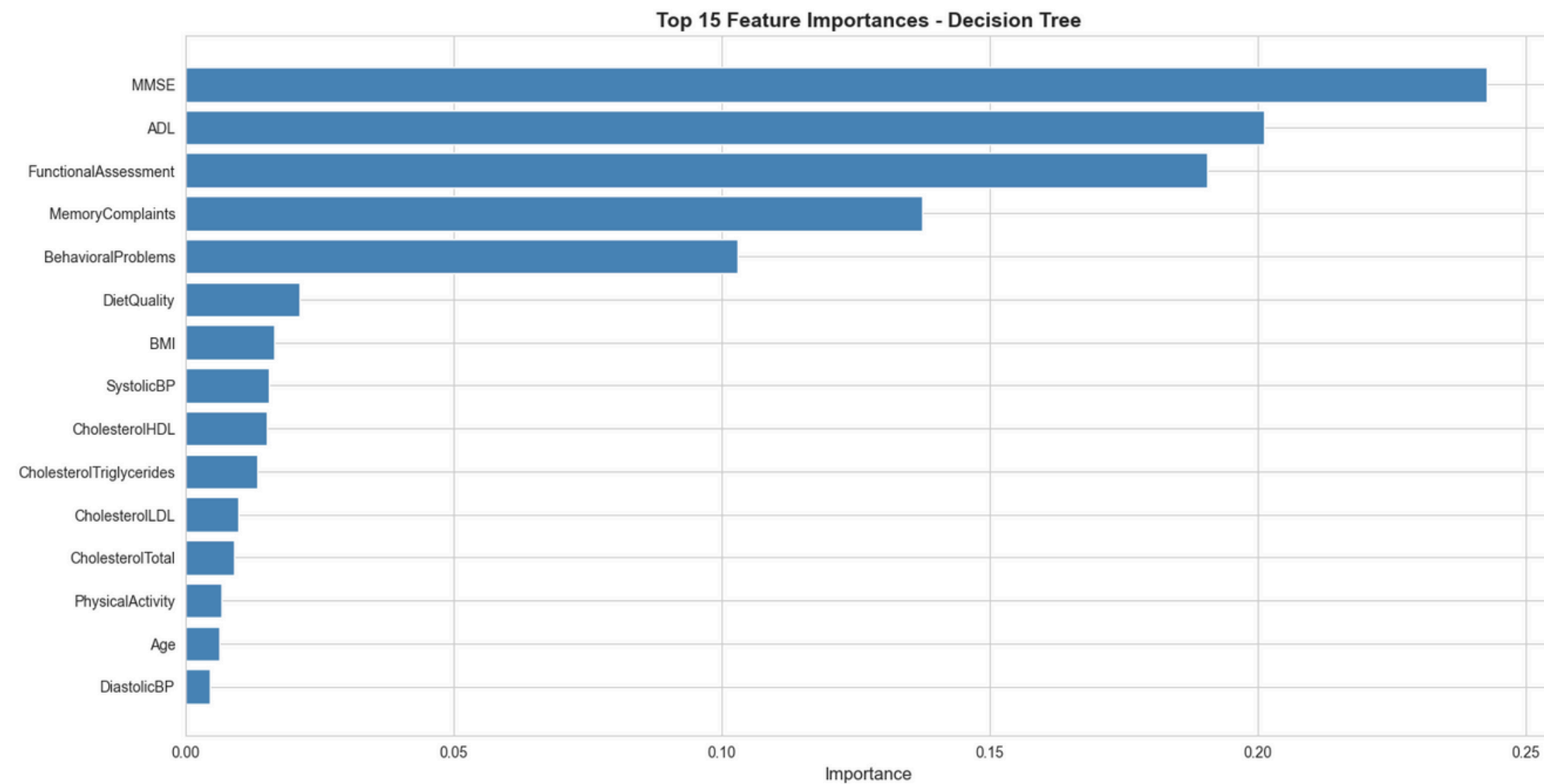
TESTING AND VALIDATION

Model	Accuracy	Precision	Recall	F1-Score	CV Accuracy	ROC-AUC
Logistic Regression	0.8140	0.8145	0.8140	0.8142	0.8395	0.886
Decison Tree	0.8884	0.8884	0.8884	0.8884	0.9069	0.878
SVM	0.8326	0.8305	0.8326	0.8302	0.8336	0.894
KNN	0.7465	0.7418	0.7465	0.7308	0.7254	0.767



FEATURE IMPORTANCE

- MMSE and ADL are most influential.
- Functional assessment, memory complaints, behavioral changes moderate impact.
- Others minor.





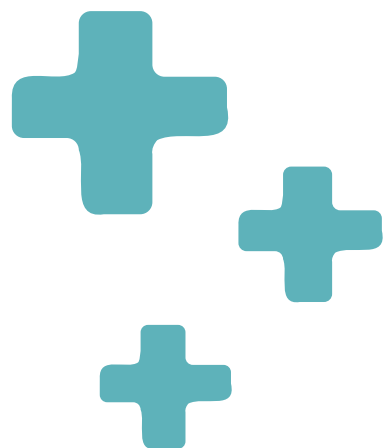
STAKEHOLDER BENEFITS

Healthcare Professionals

- Decision-support tool for early screening
- Faster and more consistent risk assessment
- Supports clinical judgment, not replaces it

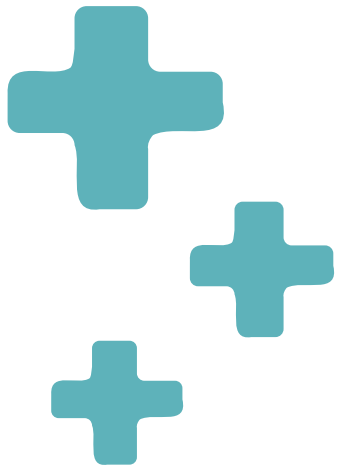
Patients

- Earlier risk detection and intervention
- Better planning for treatment and lifestyle changes
- Reduced anxiety through clearer risk assessment



CONCLUSION

- Decision Tree classifier is the most effective model for predicting Alzheimer's in this dataset.
- MMSE and ADL are key predictors.
- Machine learning models can support early AD detection and assist clinical decision-making.
- Future work: incorporate more data, advanced models (ensemble/deep learning), feature optimization.



THANK YOU !

